AutoScout24 Car Dataset

In this project, I am analyzing a car dataset obtained from www.autoscout24.de, which is a platform in Germany for buying and selling cars. The dataset consists of 10 columns and 46,406 rows, where each row represents a distinct transaction.

To facilitate the analysis, I have divided it into six parts. The columns in the dataset include car mileage, make, model, fuel type, offer type, gear type, sales price, horsepower, and year of make. Additionally, I have created an additional column called "car" by combining the make, model, and year of make to streamline the analysis process.

In the first part, I will extract insights regarding the car makers, exploring information related to different car manufacturers. The second part will focus on examining the various types of cars present in the dataset.

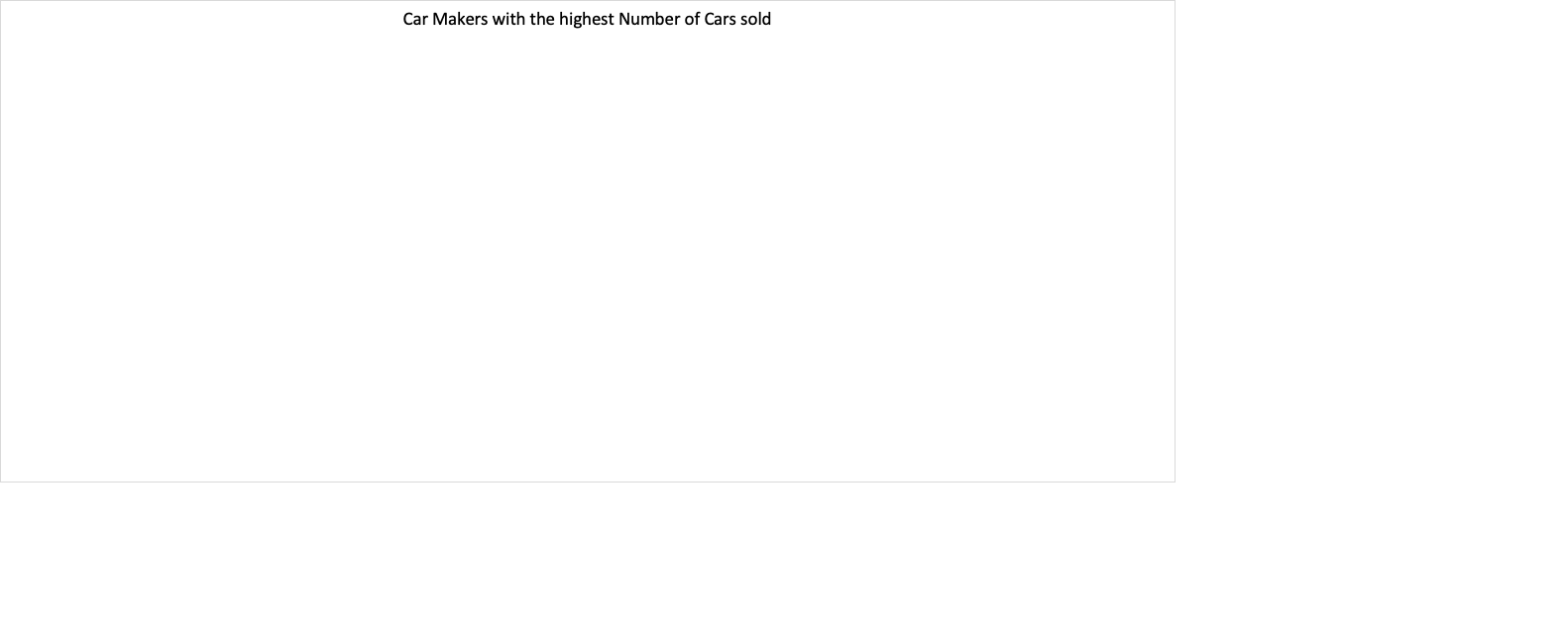
Moving on to the third part, I will delve into the analysis of different fuel types, offer types, and gear types represented in the dataset. This will provide a better understanding of the distribution of these variables.

In the fourth part, I will analyze the distribution of numerical variables present in the dataset, gaining insights into their ranges and variations.

Part five will explore the relationships between the price, mileage, horsepower, and year of make. This analysis aims to uncover any potential correlations and patterns among these variables.

Lastly, in the final part, I will construct both a multiple linear regression model and a simple linear regression model. These models will help evaluate how much of the variation in price (the dependent variable) can be explained by the independent variables, such as mileage, horsepower, and year of make. This will provide valuable insights into the factors influencing car prices in the dataset.

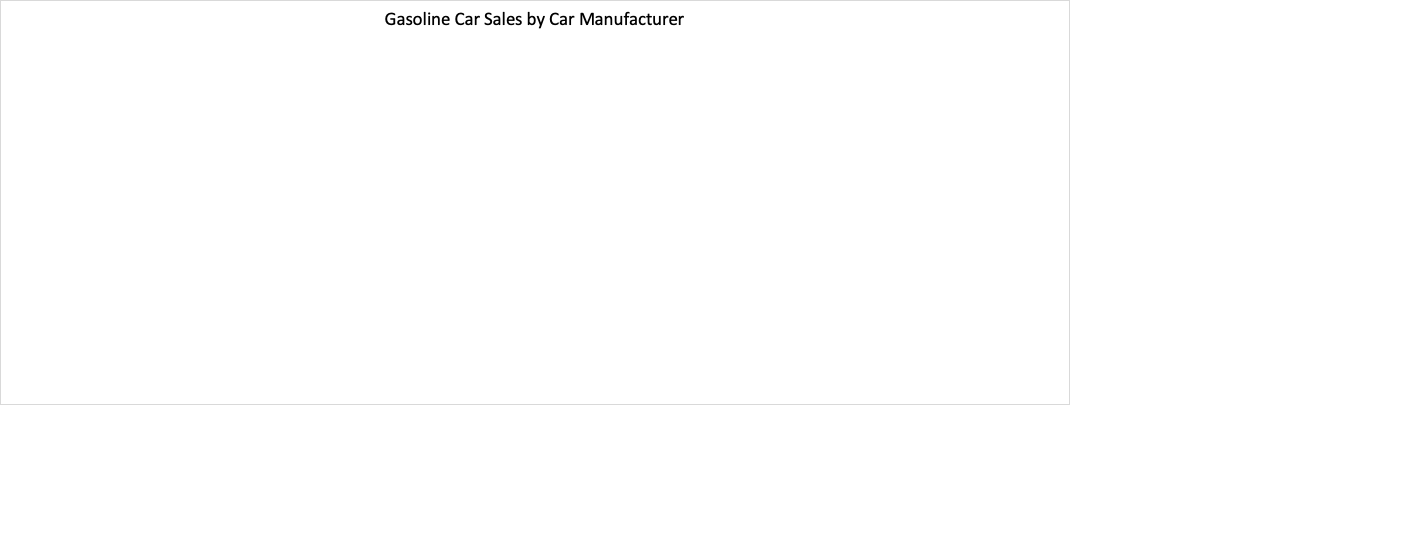
Part I Car Manufacturers



In our dataset, Volkswagen emerges as the leading car manufacturer with the highest number of cars sold, totaling 6,931 vehicles. Following Volkswagen, Opel secures the second position with 4,814 cars, while Ford ranks third with 4,442 cars.

The graph displays the car makers with the highest average sales prices on autoscout24 are displayed. The average sales price of Maybach cars stands out at approximately 450,000 Euros, making it the most expensive car brand by a significant margin. Ferrari secures the second position with an average sales price of around 324,000 Euros, followed by Lamborghini in third place with an average price of 305,000 Euros. These figures highlight the premium pricing associated with these luxury car brands on the autoscout24 platform.

Among all car manufacturers, Volkswagen stands out as the top seller of diesel-powered trucks and cars, with a total of 2,873 units sold. Ford takes the second position with 1,641 units sold, closely followed by Audi in third place with 1,478 units sold. These figures indicate the strong market presence of Volkswagen in the diesel truck segment, with Ford and Audi also showing notable sales numbers in this category.



Volkswagen stands out as the leading car manufacturer in our dataset, having sold the greatest number of both gasoline and diesel-powered cars. For gasoline cars specifically, Volkswagen sold 3,911 units, followed by Opel with 3,264 units, and Ford with 2,673 units.

In terms of average price, Volkswagen falls in the middle range with an average price of around 16,000 Euros. However, it is worth noting that Volkswagen does not rank in the top 15 most affordable cars according to another analysis. This suggests that the popularity of Volkswagen among German consumers may be attributed to a combination of factors such as price, quality, brand reputation, and other features.

Considering the high sales numbers across different fuel types and the popularity among consumers, it is plausible to speculate that the balance between price and quality plays a significant role in Volkswagen's success within the German market.

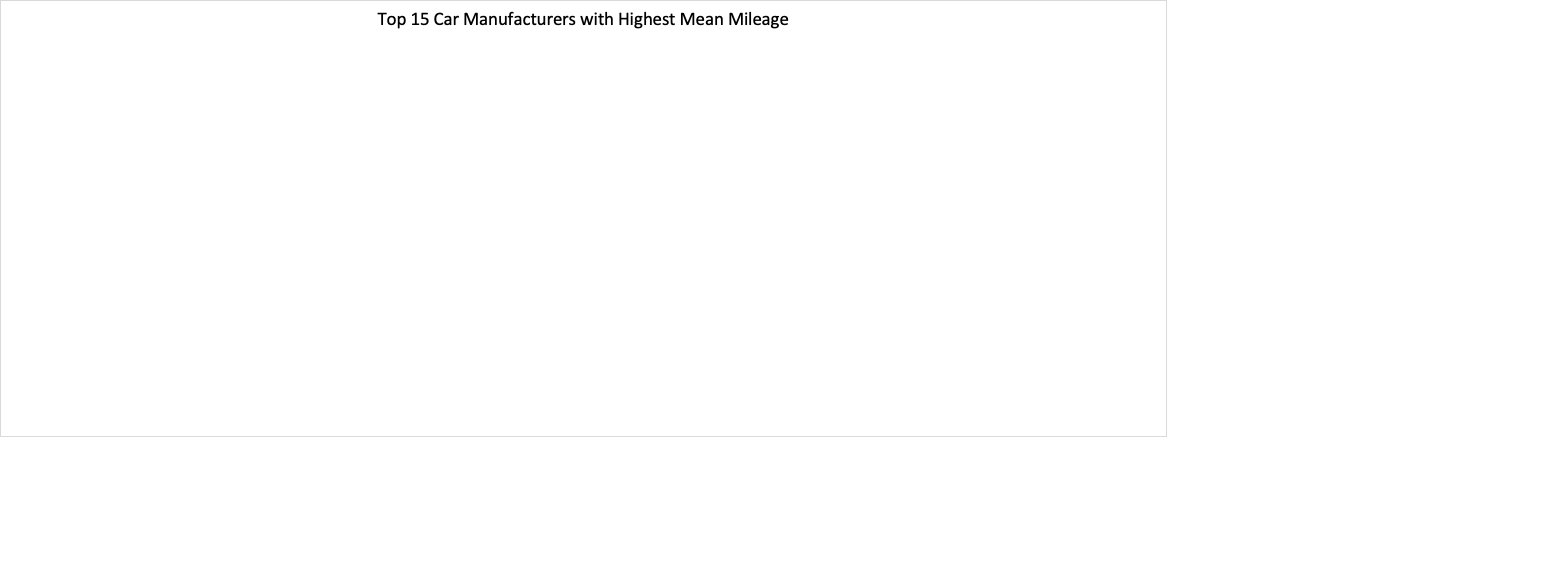
Renault emerges as the leading car manufacturer in terms of electric car sales within the dataset, having sold the greatest number of units with 299. Following Renault, smart takes the second position with 101 units, and Volkswagen ranks third with 65 units sold.

Interestingly, despite Tesla's reputation as one of the most successful electric car companies globally, it does not feature in the top 5 of this particular dataset. One possible explanation could be that Tesla primarily sells its cars through its own website, with many consumers placing orders online rather than through platforms like autoscout24.de.

Apart from Renault, smart, and Volkswagen, other notable car manufacturers in the electric car segment include Nissan, Opel, Tesla, BMW, Porsche, Audi, and Mazda, among others. These figures indicate the growing presence and popularity of electric cars from various manufacturers in the market.

In our dataset, Toyota takes the lead as the car manufacturer with the highest number of hybrid cars sold, totaling 293 units. Fiat secures the second position with 164 units sold, followed by Ford in third place with 110 units. These figures highlight the strong market presence of Toyota in the hybrid car segment, with Fiat and Ford also showing notable sales numbers in this category. Hybrid cars continue to gain popularity among consumers due to their fuel efficiency and reduced environmental impact.

According to the bar graph, the car brands with the lowest average car prices are displayed. Among the brands listed, Brilliance stands out with an average sales price of 1500 Euros. It is worth noting that Brilliance is relatively unknown compared to more prominent car brands. Additionally, the graph mentions other lesser-known brands such as Iveco and 9ff, which likely contribute to the overall low average prices in this category. These lower-priced car brands may appeal to budget-conscious buyers or those seeking affordable transportation options.



According to the data, Trucks-Lkw cars had the highest average mileage when sold, reaching 312,000 miles. Following closely behind is Iveco cars with an average mileage of 271,000 miles. In third place, an unknown car company named DAF stands out with an average mileage of 183,000 miles. It's important to note that these three car companies specialize in the sale of large trucks, which typically have higher mileage due to their usage in commercial and heavy-duty applications. The high average mileage reflects the durability and robustness of these trucks, as they are designed to handle extensive mileage over their lifetime.

Part II Cars

The graph highlights the cars with the highest number of units sold on autoscout24. Taking the top spot is the Fiat 500 2021, with an impressive 232 units sold. Following closely behind is the Ford Fiesta 2019, which sold 204 units.

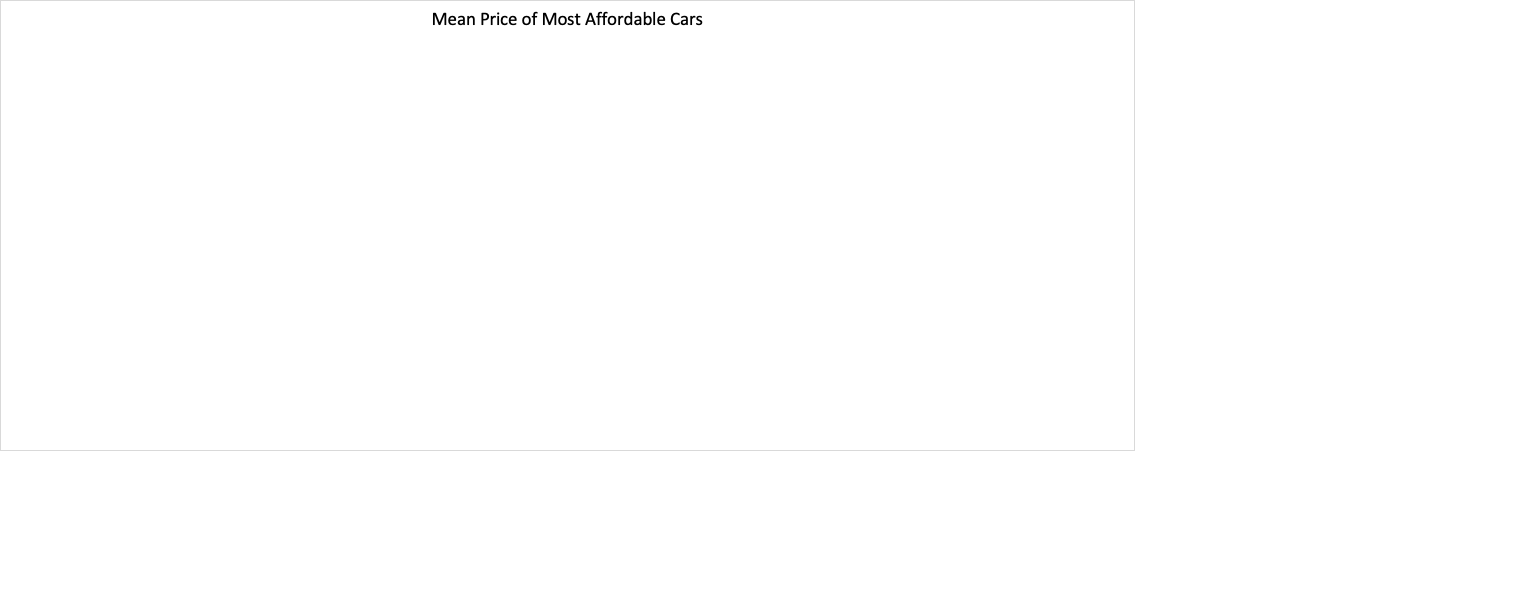
Notably, Volkswagen emerges as a popular choice among buyers, as seven different Volkswagen models make it to the top 15 best-selling cars in the dataset. This signifies the strong market presence and consumer demand for Volkswagen vehicles. The popularity of these models may be attributed to factors such as brand reputation, affordability, and performance.

According to the data, the Ferrari F12 2017 holds the distinction of being the most expensive car sold on the platform, with a sale price of approximately 1.2 million Euros. This high price reflects the exclusivity and desirability of the Ferrari brand, as well as the exceptional performance and luxury offered by the F12 model.

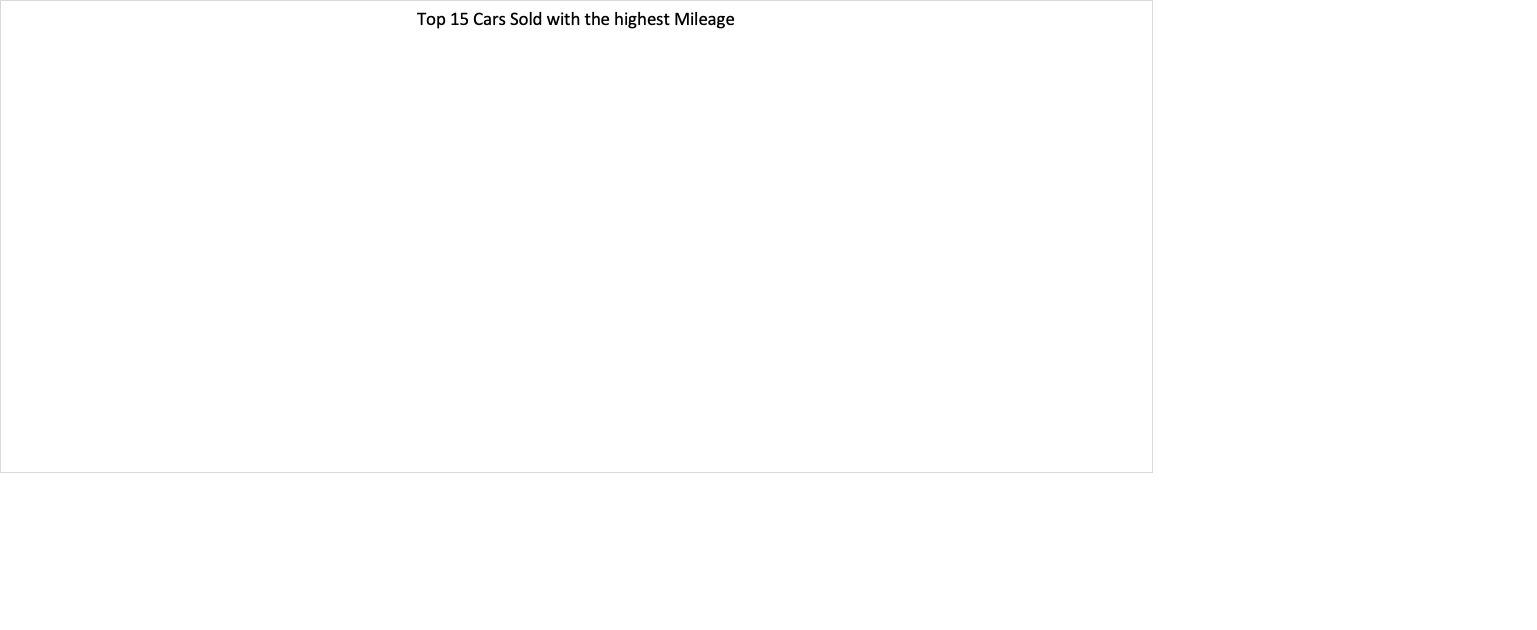
In second place, we have the Maybach Pullman, commanding a price of around 717,000 Euros. Maybach is renowned for its ultra-luxurious and spacious vehicles, and the Pullman is no exception. The high price tag reflects the premium features, craftsmanship, and prestige associated with owning a Maybach Pullman.

These exceptionally high-priced cars cater to a niche market of luxury car enthusiasts who seek unparalleled performance, craftsmanship, and status symbol in their automotive choices.

According to the provided information, the Mercedes-Benz G 63 AMG 2019 is hailed as the fastest car in terms of horsepower, boasting an impressive 850 horsepower. In terms of horsepower, both the Ferrari 812 2021 and the Audi R8 2013 are tied for second place, with a notable horsepower rating of 799. These high-performance vehicles from renowned brands deliver remarkable speed and thrilling performance on the road.



The graph illustrates the top 15 cheapest cars sold on autoscout24. The Lada Priora 2011 takes the crown as the least expensive car, with a remarkably low sale price of 1190 Euros. Following closely behind is the Brilliance BS4, which was sold for 1500 Euros, making it the second most affordable option on the list.



The car with the highest recorded mileage in our dataset is the Opel Karl 2019, which astonishingly reached over 1.1 million miles before being sold. Following closely behind is a BMW 320 2014, which had accumulated nearly 1 million miles at the time of sale.

Additionally, among the cars listed with high mileage, we observe the presence of five different Mercedes-Benz models. This indicates the durability and longevity of Mercedes-Benz vehicles, as they can withstand substantial mileage over extended periods of use.

Part 3 Categorical Variables

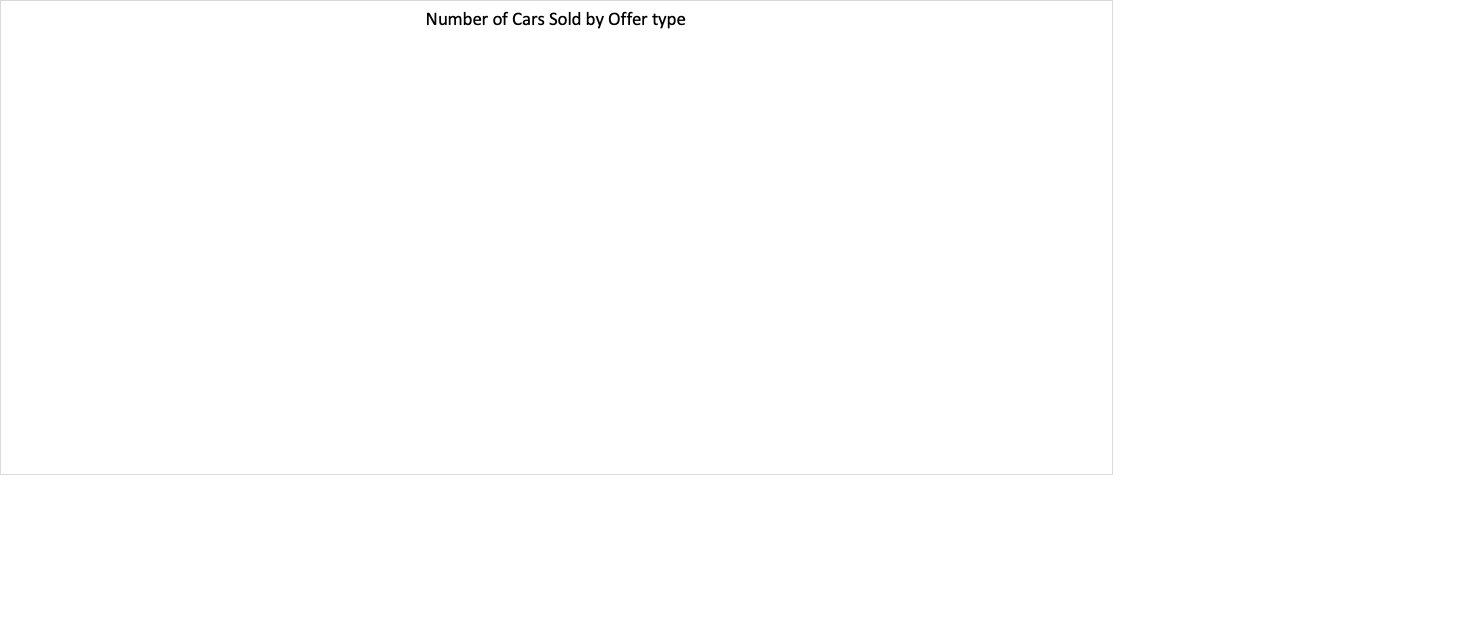
In our dataset, we have observed the sales of different fuel types among the cars. Gasoline cars recorded the highest number of units sold, with a total of 28,864 vehicles. 15,222 Diesel Cars were sold. Hybrid cars ranked third in terms of sales, with 1,158 units sold. Lastly, electric cars occupied the fourth position with 756 units sold.

These figures highlight the popularity and demand for gasoline-powered vehicles, as they constitute a significant majority among the various fuel types. Diesel cars also maintain a strong presence in the market, attracting a substantial number of buyers. Additionally, the sales of hybrid and electric cars, while comparatively lower, indicate a growing interest in more environmentally-friendly and energy-efficient vehicles.

It's worth noting that these sales figures are specific to our dataset and may not reflect the overall market trends. The preferences and choices of car buyers can vary based on factors such as availability, pricing, and individual preferences.

The bar graph displays the mean car price for every fuel type. Electric/Diesel cars commanded the highest average at 37605 Euros.

In our dataset, there was one Hydrogen car that sold for 34,990 Euros. Hybrid cars (Electric/Gasoline) sold for an average price 29,695 Euros. Electric cars sold for 23,052 Euros. Diesel cars were slightly costlier than Gasoline car at 18,122 Euros. Gasoline cars sold for 15,074 Euros.



In our dataset, used cars dominate as the most prevalent offer type, with a total of 40,121 cars. These are cars that have previously been owned by individuals and are now being sold on the market.

Following used cars, pre-registered cars make up the second-highest offer type, with a count of 2,780 cars. Pre-registered cars are vehicles that have been registered by dealerships but have not yet been sold to customers. These cars are often used for promotional purposes or to meet sales targets set by manufacturers.

Demonstration cars rank third in terms of offer type, with a total of 2,368 cars. These are vehicles that are used by dealerships for test drives and showcasing to potential customers. They typically have low mileage and are well-maintained.

Lastly, we have 1,121 cars provided by companies to their employees as a perk. These employee cars are often offered as part of employment packages or incentives, allowing employees to use them for personal purposes.

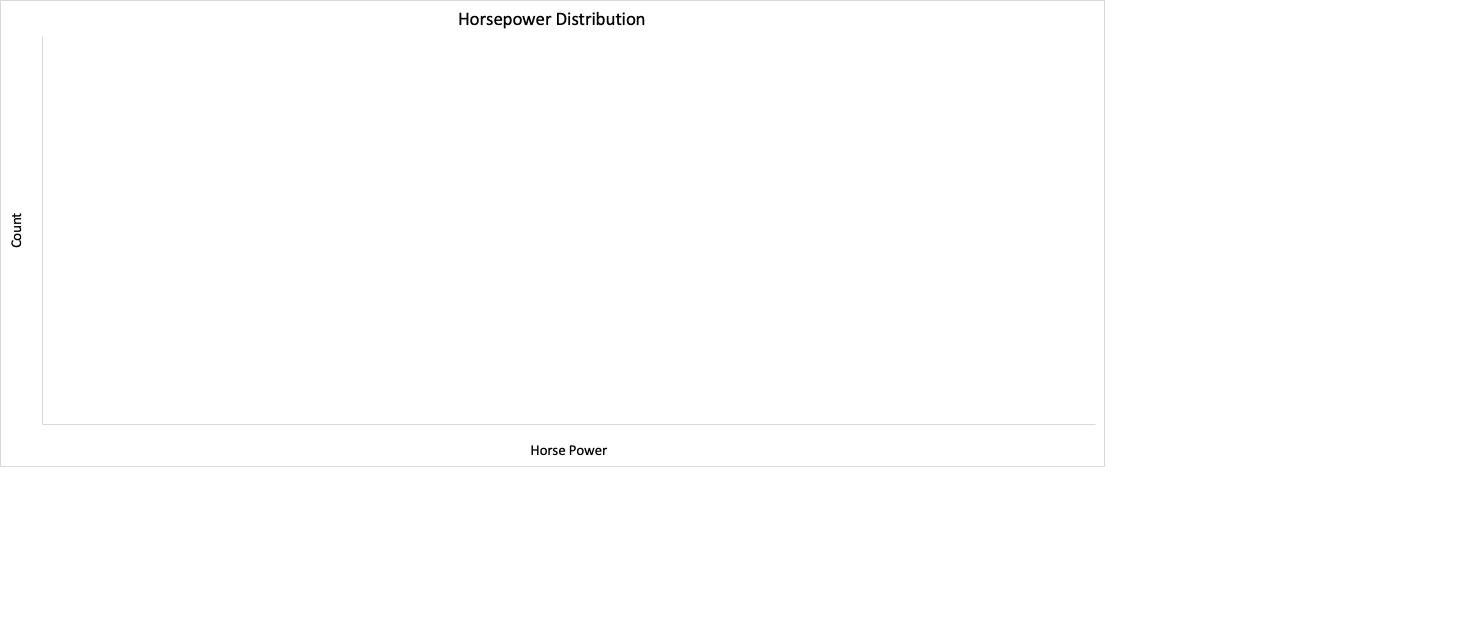
The bar chart presents the average sales prices for different offer types of cars. New cars are the most expensive among the listed categories, with an average sale price of 66,000 Euros. Demonstration cars, which are used for test drives and showcasing purposes, have the second highest average sales price of 34,960 Euros. These cars are often well-maintained and have relatively low mileage, contributing to their higher price compared to other categories. Employee cars, provided by companies to their employees as perks, have an average sale price of 30,000 Euros. Used cars, on the other hand, have the most affordable average sales price of 14,700 Euros. As previously owned vehicles, they generally come at lower prices compared to new or demonstration cars.

The graph illustrates the unit sales of cars based on their gear type. Manual cars are the most prevalent in the dataset, with a total of 30,380 units sold. This indicates that there is still a significant demand for manual transmission cars, particularly in the German market. Automatic cars are the second most popular gear type, with 15,786 units sold. There are 56 units of semi-automatic cars sold, which represent a smaller portion of the total sales. Semi-automatic transmissions combine features of both manual and automatic transmissions, providing drivers with the option to shift gears manually or let the system handle it automatically.



According to the data, the average sale price of an automatic car is 28K Euros, which is higher compared to the average sale price of manual cars at 10K Euros and semi-automatic cars at 8,425 Euros. This suggests that automatic cars tend to be priced higher than their manual and semi-automatic counterparts.

Part IV Distribution of Numerical variables



|  |  |
| --- | --- |
| Mean | 133.0 |
| Standard Error | 0.4 |
| Median | 116.0 |
| Mode | 150.0 |
| Standard Deviation | 75.4 |
| Sample Variance | 5692.6 |
| Kurtosis | 12.3 |
| Skewness | 2.8 |
| Range | 849.0 |
| Minimum | 1.0 |
| Maximum | 850.0 |
| Sum | 6167590.0 |
| Count | 46376.0 |

The histogram shows the distribution of horsepower in the dataset. The shape of the distribution is slightly skewed to the right, indicating that there are cars with higher horsepower than the average.

The average horsepower of the cars in the dataset is 133. This means that, on average, the cars have a horsepower of 133.

The median, which is the middle value in the dataset, is 116.

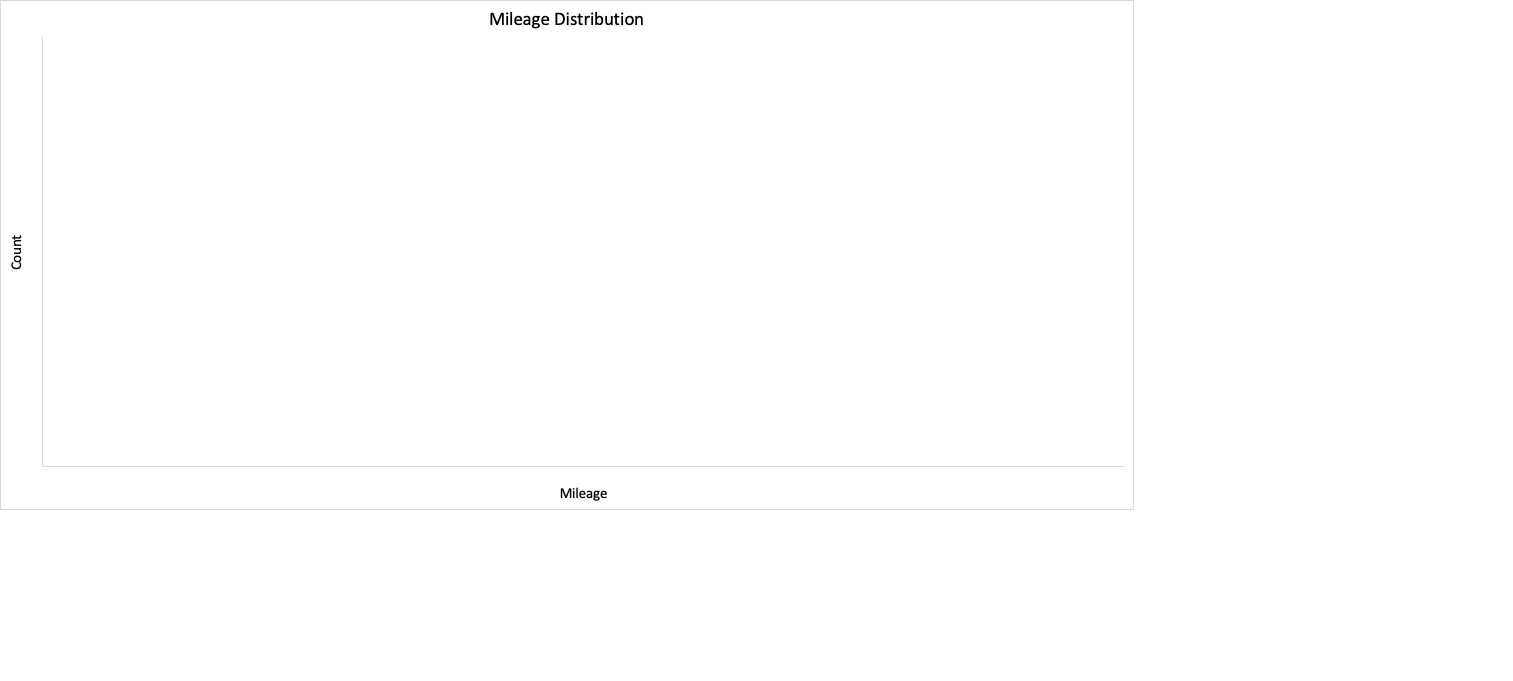
The mode, which is the most frequent value in the dataset, is 150. This indicates that there are more cars with a horsepower of 150 than any other specific value.

The skewness of 2.8 indicates a right skew, meaning that the distribution has a tail on the right side. This further supports the presence of cars with higher horsepower in the dataset.

The kurtosis of 12.3 indicates that the distribution has a high peak and heavy tails, suggesting the presence of outliers.

The range of the horsepower values in the dataset is 849, with the fastest car having 850 horsepower and the slowest car having 1 horsepower.

The standard deviation of 75.4 indicates a significant spread and variation in the dataset. This means that the horsepower values are relatively dispersed from the average.

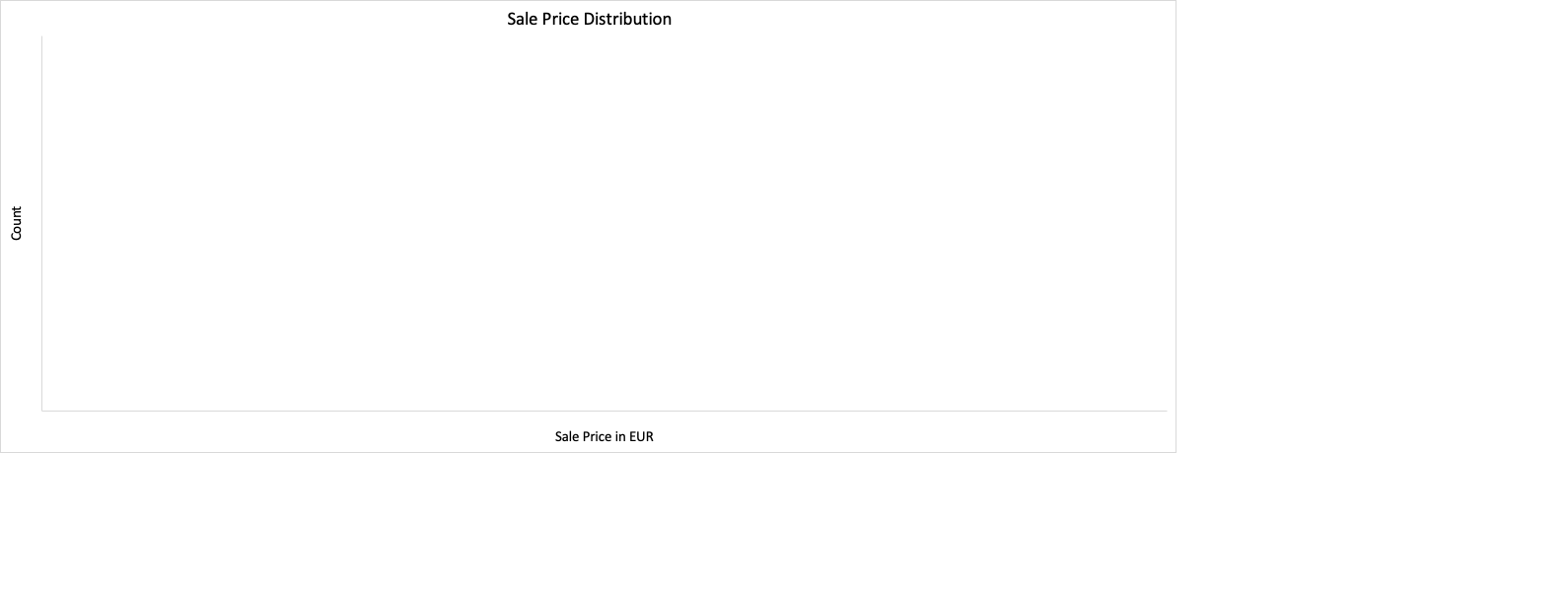


|  |  |
| --- | --- |
| *mileage* | |
|  |  |
| Mean | 71177.9 |
| Standard Error | 290.7 |
| Median | 60000.0 |
| Mode | 10.0 |
| Standard Deviation | 62625.3 |
| Sample Variance | 3921929259.3 |
| Kurtosis | 5.6 |
| Skewness | 1.4 |
| Range | 1111111.0 |
| Minimum | 0.0 |
| Maximum | 1111111.0 |
| Sum | 3303008784.0 |
| Count | 46405.0 |

The histogram displays the distribution of mileage in the dataset. The skewness of 1.4 indicates a right-skewed distribution, which means there are cars or trucks with higher mileage than the average.

The kurtosis of 5.6 suggests that the distribution has heavy tails, indicating the presence of outliers on the higher end of the mileage distribution. The average mileage of the cars in the dataset is 71.11K miles and the median, which is the middle value in the dataset, is 60K miles. The mode, which is the most frequent value in the dataset, is 10K miles.

The standard deviation of 62262 indicates significant variability in the dataset. This means that the mileage values are dispersed from the average. The range of the mileage values in the dataset is 1.1 million miles, with the truck with the highest mileage having 1.1 million miles and some of the newer cars having 0 miles.



|  |  |
| --- | --- |
| *price* | |
|  |  |
| Mean | 16572.3 |
| Standard Error | 89.6 |
| Median | 10999.0 |
| Mode | 7990.0 |
| Standard Deviation | 19304.7 |
| Sample Variance | 372671284.7 |
| Kurtosis | 450.8 |
| Skewness | 12.5 |
| Range | 1198800.0 |
| Minimum | 1100.0 |
| Maximum | 1199900.0 |
| Sum | 769039309.0 |
| Count | 46405.0 |

The histogram displays the distribution of sale prices in the dataset. The average sales price is 16.5K Euros and the median, which is the middle value in the dataset, is 10.99K Euros. The skewness of 12.5 indicates positive skewness, meaning that the distribution is skewed to the right. This indicates that there are outliers or cars with significantly higher sale prices than the average. The kurtosis of 450.8 suggests a relatively peaked distribution with heavy tails, indicating many outliers with extremely high sale prices. The range of sale prices in the dataset is 1.198 million Euros, with the highest-priced car being significantly more expensive than the average car. The standard deviation of 19304 indicates a great variability in sale prices, meaning that the sale prices are spread out from the average.

Part V Relationships between Numerical Variables

The scatterplot shows the relationship between the mileage of a car and its sales price. The correlation coefficient of -0.30 indicates a moderate negative relationship between the two variables. This means that as the mileage of a car increases, the sales price tends to decrease. However, it is important to note that the relationship is not very strong, as indicated by the moderate correlation coefficient.

The scatterplot displays the relationship between the year of make and the sales price of cars. The correlation coefficient of 0.4 indicates a moderate positive relationship between the two variables.

This means that there is a tendency for newer cars to have higher sales prices. As the year of make increases, the sales price tends to increase as well. However, it's important to note that the relationship is not very strong, as indicated by the moderate correlation coefficient.

The scatterplot displays the relationship between horsepower and sales price of cars. The correlation coefficient of 0.75 indicates a strong positive relationship between the two variables.

This means that there is a clear tendency for cars with higher horsepower to have higher sales prices. As the horsepower increases, the sales price tends to increase as well. The strong positive correlation suggests that horsepower is a significant factor in determining the sales price of a car.

Part VI Prediction of Sales Price

For this part, I created a simple regression model to predict the sales price. The independent variable is horsepower, and the dependant variable is the sales price.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.74715771 |  |  |  |  |  |  |  |
| R Square | 0.55824464 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.55823512 |  |  |  |  |  |  |  |
| Standard Error | 12831.1139 |  |  |  |  |  |  |  |
| Observations | 46403 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 9.65381E+12 | 9.65381E+12 | 58636.7743 | 0 |  |  |  |
| Residual | 46401 | 7.63934E+12 | 164637484 |  |  |  |  |  |
| Total | 46402 | 1.72932E+13 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -8841.7067 | 120.6783282 | -73.2667319 | 0 | -9078.2381 | -8605.1754 | -9078.2381 | -8605.1754 |
| hp | 191.109964 | 0.789220386 | 242.1503135 | 0 | 189.56308 | 192.656848 | 189.56308 | 192.656848 |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

The R-squared value of 0.55 indicates that 55% of the variation in sales price can be explained by the horsepower variable in your regression model. This means that horsepower has a moderate level of predictive power for determining the sales price of cars.

The p-value of 0 suggests that the horsepower variable is statistically significant in predicting the sales price. This means that there is evidence to support the idea that horsepower is a useful predictor of sales price.

The t-statistic being significant further confirms the statistical significance of the horsepower variable in the model. It indicates that the estimated coefficient for horsepower is unlikely to be zero.

The sum of squares regression being larger than the sum of squares residual suggests that the regression model captures a significant portion of the variation in sales prices. This implies that the model is doing a good job of explaining the relationship between horsepower and sales price.

However, it's important to note that while the model is statistically significant and explains a reasonable amount of the variation in sales price, there may be other factors that influence the sales price of cars that are not accounted for in this model.

The intercept value of -8,841 represents the expected sales price when the horsepower value is zero. In this case, it doesn't make practical sense because cars cannot have negative sales prices. This highlights the limitation of the model when extrapolating beyond the observed range of data.

Overall, while the model provides some insights into the relationship between horsepower and sales price, it is important to consider other factors and potentially include them in the model to enhance its predictive accuracy.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.81060706 |  |  |  |  |  |  |  |
| R Square | 0.65708381 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.65706164 |  |  |  |  |  |  |  |
| Standard Error | 11305.1622 |  |  |  |  |  |  |  |
| Observations | 46403 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 3 | 1.13631E+13 | 3.78768E+12 | 29636.0384 | 0 |  |  |  |
| Residual | 46399 | 5.9301E+12 | 127806692.8 |  |  |  |  |  |
| Total | 46402 | 1.72932E+13 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -2011573.6 | 46617.04359 | -43.15103285 | 0 | -2102943.7 | -1920203.5 | -2102943.7 | -1920203.5 |
| hp | 183.463543 | 0.71177465 | 257.755096 | 0 | 182.068454 | 184.858632 | 182.068454 | 184.858632 |
| year | 995.898599 | 23.10561366 | 43.10201899 | 0 | 950.611247 | 1041.18595 | 950.611247 | 1041.18595 |
| mileage | -0.0561333 | 0.001147903 | -48.90072909 | 0 | -0.0583832 | -0.0538834 | -0.0583832 | -0.0538834 |

It's great to see that you have expanded your regression model by including two additional variables, year of make and mileage. The improved R-squared value of 0.65 indicates that the model now explains 65% of the variation in sales prices, which is an improvement from the previous model.

The multiple coefficients estimate of 0.81 suggests a strong positive relationship between the independent variables (horsepower, year of make, and mileage) and the dependent variable (sales price). This means that these variables collectively have a significant impact on predicting the sales price of cars.

The p-value being zero for all independent variables indicates that all the variables in the model are statistically significant and useful in predicting the sales price. This provides confidence in the reliability of the estimated coefficients.

The intercept coefficient of -2,011,573 represents the expected sales price when all independent variables are zero, which is not a meaningful interpretation in this context. It's important to note that the intercept value may not have practical significance when the independent variables are not within the observed range of the dataset.

The coefficient estimates of horsepower being 183 suggests that for every one unit increase in horsepower, the sales price is expected to increase by 183 Euros, holding other variables constant. Similarly, the coefficient estimate of the year of make being 995 indicates that for every one unit increase in the year of make, the sales price is expected to increase by 995 Euros, assuming other variables remain the same. On the other hand, the coefficient estimate of mileage being -0.056 suggests that for every one unit increase in mileage, the sales price is expected to decrease by 0.056 Euros, assuming other variables are constant.

Overall, this expanded model with horsepower, year of make, and mileage as independent variables provides a better explanation of the sales price variation. However, it's important to note that there may still be other factors influencing the sales price that are not accounted for in this model.

A car with a mileage of 71177 miles, 132 hp and year of make of 2011 would be worth around 9954 Euros.

Sales Price = -2011573.6 + 183.463543 \* 132 + 995.898599 \* 2011 -0.0561333\*71177